コンピュータ診断支援装置の性能評価

開発ガイドライン２０１５

（手引き）

平成２７年１２月

経済産業省／国立研究開発法人日本医療研究開発機構

目次

APPENDIX

コンピュータ診断支援装置の性能評価 開発ガイドライン 2015（手引き）

1. 本ガイドラインの目的

コンピュータ検出／診断支援（Computer-Aided Detection/Diagnosis; CAD）とは、X 線画像に代表される放射線画像をはじめとする医用画像をコンピュータで定量的に解析し、「医師による診断」を側面より支援する行為であり、またその機能を提供するソフトウェアあるいはそれを具備する装置を指す。本ガイドラインは、医療機器としての CAD 装置の研究開発と薬事法に関わる承認申請（薬事申請）を円滑化・加速することを目的として策定されたものであり、「コンピュータ診断支援装置に関する評価指標」（平成 23 年 12 月 7 日薬食機発 1207 第 1 号（別添 3））および「コンピュータ診断支援装置におけるソフトウェア設計・開発管理」（医療機器開発ガイドライン：経済産業省、平成 25 年 3 月）とともに活用されることを想定して編集した。

2. 定義と用途

コンピュータ検出／診断支援（Computer-Aided Detection/Diagnosis; CAD）は目的に応じて２つに大別され、解析結果に基づいて病変候補位置の情報をマーカで医師に示すだけのものはコンピュータ検出支援（Computer-Aided Detection; CADe）、質的診断に関する情報までも提示するものはコンピュータ診断支援（Computer-Aided Diagnosis; CADx）と呼ばれる（詳細は Appendix「A.1 CAD の分類と CADx の位置付け」参照）。本ガイドラインが対象とするのは後者の CADx であり、その定義や用途を以下に示す：

定義： コンピュータ診断支援ソフトウェア [CADx (Computer-Aided Diagnosis)]
画像を解析して内蔵する基準に基づいて病変の候補部位をコンピュータが自動的に分析し、医学的に広く臨床で用いられている診断基準に基づく質的診断に関する情報を提供するソフトウェア、あるいはそれを具備する装置。

用途： 本ガイドラインが対象とするコンピュータ診断支援ソフトウェアは、医師の診断を支援するために用いられるものである。
最終診断結果には医師が責任を負う必要があるため、コンピュータ診断支援ソフトウェアの出力結果のみに判断を委ねてはならない。

3. 適用範囲

本ガイドラインでは、質的診断に関する情報として、良悪性などに関するクラス分類の結果を出力する場合を対象とする。従って、クラス分類結果を出力しないもの（たとえば、

腫瘍のサイズなどの特徴量、単一あるいは複数の特徴量から導かれた悪性度や進行度などの連続値、類似症例の検索結果などを出力するだけのもの）は本ガイドラインの対象には含めない。本ガイドラインの適用を受けるためには、必ずクラス分類結果を出力し、それに対して適切に評価を行わなければならない（Appendix 「A.2 性能評価の基礎」参照）。その他、再学習により性能が自動的に変化する機能を有する CADx については、現時点では本ガイドラインの対象には含めない。

4. 評価手法および留意事項

本節では、開発段階で実施すべき「安全性と品質管理」および「性能評価」に関する試験の設定根拠について説明する。

4.1 安全性と品質管理

CADx はソフトウェア単体の場合とハードウェアを含む場合があるが、いずれについても、安全性・品質管理に関する JIS 規格などの基準が存在する（表１）。

	ソフトウェアが搭載されたハードウェア		ソフトウェア単体
	非患者環境下	患者環境下	
電気的安全性	JIS C 6950-1 EMC（JIS T 0601-1-2 または CISPR 24 (immunity), CISPR 22 (emission)）	JIS C 6950-1 JIS T 0601-1 16 章 ME システム EMC（JIS T 0601-1-2 または CISPR 24 (immunity), CISPR 22 (emission)）	該当なし
品質マネジメント	QMS 省令	QMS 省令	QMS 省令
リスクマネジメント	JIS T 14971	JIS T 14971	JIS T 14971
ソフトウェアの品質管理			「コンピュータ診断支援装置におけるソフトウェア設計・開発管理 開発ガイドライン 2012」が活用できる

表 1　CADx の安全性・品質管理に関連する基準例

（ア）基本的考え方

CADx の主体はソフトウェアにて規定される性能であり、ハードウェアに搭載して利用される。そのため、安全性に関しては、汎用画像診断装置ワークステーションと扱いと類似する。患者環境外（患者に接しない環境）で使用される汎用ワークステーションに対しては、JIS C 6950-1 が存在する。患者環境内（患者に接する可能性がある環境）で使用される画像診断ワークステーションに対しては、それに加えて JIS T 0601-1（医用電気機器—第１部：基礎安全及び基本性能に関する一般要求事項：第 16 章 ME システム）が存在する。医療機器の品質マネジメントについては、QMS 省令への遵守が求められてい

る。（Appendix「A.3 QMS (Quality Management System)」参照）

（イ）推奨事項

医療機器としての法規上の要求事項としては、QMS 省令に基づき品質マネジメントシステムを構築することが求められている。一方で、開発時から考慮すべき規格やガイドラインとして、リスクマネジメントには JIS T 14971 が、医療機器のソフトウェアのライフサイクルプロセスには、JIS T 2304（医療機器ソフトウェア―ソフトウェアライフサイクルプロセス）、IEC/TR80002-1（Medical device software - Part 1: Guidance on the application of ISO 14971 to medical device software; 医療機器ソフトウェアー第 1 部：医療機器ソフトウェアへの ISO14971 の適用の手引き）、「コンピュータ診断支援装置におけるソフトウェア設計・開発管理 開発ガイドライン 2012」などが策定されている。また、薬事申請するときには、「コンピュータ診断支援装置に関する評価指標」（平成 23 年 12 月 7 日薬食機発 1207 第 1 号（別添 3））も同時に参照することを推奨する。

4.2 システム開発及び性能評価に用いるデータ

CADx のシステム開発（パラメータ調整や動作確認など）及び性能評価の際には、妥当な性質ならびに数量の臨床画像や、ファントムなどを用いて生成された画像を使用する。ファントムなどを用いて生成した画像で性能評価を行う際は、その画像が具備する性質が臨床画像と同質であることを科学的・論理的に説明する。

（ア）基本的考え方

性能評価を適切に行うため、CADx のシステム開発（パラメータ調整や動作確認など）に用いるデータベースと性能評価に用いるデータベースは厳格に区別する。両者を混在させない。

開発時の評価にあたっては、下記項目を明確にすることが望まれる。

(1) 機能の明確化

CADx が対象とする臓器や病状（疾患名、良悪性の判定、重症度など）を明らかにする。

(2) 撮影条件の明示

CADx が有する性能を正しく評価するために、適用できるモダリティ（検査機器の種類）、入力となる医用画像データの収集条件（撮影時の測定パラメータなど）、入力画像の仕様（経時画像、造影剤の有無）、画像の空間的な歪みや背景雑音に関して

詳細なデータを分析し、適応可能な画像の仕様（画素数、濃淡階調など）を記録することが望まれる。この結果に基づいて、開発する CADx の性能が担保できる入力画像の範囲や仕様を明示する必要がある。

(3) 正解の明示

CADx の開発時において使用した画像の正解について明示する。具体的には、病理診断結果や、合理的な診断結果（客観性、再現性）などを記述する。

(4) 評価数および感度、特異度などの明示

CADx の性能が以下のいずれかを満たすことを十分な症例数を含む適切なデータベース（Appendix「A.4 画像データベースとゴールド・スタンダード」参照）を利用して、ROC 解析および感度・特異度などを計算して明示する（Appendix「A.2 性能評価の基礎」参照）。

- 「臨床において医師が行っている診断精度」よりも、「CADx の支援した結果」が、統計的に有意に優れているか同等であること。
- 同一製品のバージョンアップなどの場合は、「既承認品の CADx の分類精度」よりも統計的に有意に優れているか同等であること。

このとき、CADx の使用による医師への影響、例えば、読影時間や疲労などについて考慮する必要がある（Appendix「A.5 性能評価のための読影実験における注意点」参照）。

（イ）推奨事項

性能評価の目的は、CADx 製品の使用目的（意図した仕様）としての診断支援が実施できることや、適切に設定したエンドポイントに従って CADx の性能を示すことにある。その際、医用画像データの取得に使用したシステムの特性および仕様を明確にするべきである。実施計画の立案、症例数、比較対照および海外データの扱いについては、「コンピュータ診断支援装置に関する評価指標」（平成 23 年 12 月 7 日薬食機発 1207 第 1 号（別添 3））を参考にする。

5. 関連資料（法令、通知、ガイドライン）

コンピュータ診断支援装置に関連する関係法令および関連通知などを以下に示す。

- JIS C 6950-1:2012「情報技術機器—安全性—第１部：一般要求事項」
- JIS T 0601-1:2014「医用電気機器—第１部:基礎安全及び基本性能に関する一般要求事項」

- JIS T 2304:2012「医療機器ソフトウェア—ソフトウェアライフサイクルプロセス」
- JIS Q 13485:2005「医療機器－品質マネジメントシステム－規制目的のための要求事項」
- JIS T 14971:2012「医療機器－リスクマネジメントの医療機器への適用」
- IEC 62304, Medical device software - Software life cycle processes（医療機器ソフトウェア—ソフトウェアライフサイクルプロセス）
- IEC/TR80002-1 （Medical device software - Part 1: Guidance on the application of ISO 14971 to medical device software; 医療機器ソフトウェア－第 1 部：医療機器ソフトウェアへの ISO14971 の適用の手引き）
- ISO 9001, Quality Management System（品質マネジメントシステム）
- CISPR 22　情報技術機器の無線妨害特性の限度値および測定法
- CISPR 24　情報技術機器のイミュニティ特性の限度値および測定法
- QMS 省令 :「医療機器及び体外診断用医薬品の製造管理及び品質管理の基準に関する省令（平成 16 年 12 月 17 日厚生労働省令第 169 号）
- GCP 省令 :「医療機器の臨床試験の実施基準に関する省令」（平成 17 年 3 月 23 日厚生労働省令第 36 号、最終改正：平成 21 年 3 月 31 日厚生労働省令第 68 号）
- GVP 省令 :「医薬品、医薬部外品、化粧品、医療機器及び再生医療等製品の製造販売後の安全管理の基準に関する省令」（平成 16 年 9 月 22 日厚生労働省令第 135 号）
- 医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律第 23 条第 2 項 5 の 3（医療機器 GCP 省令の対象となる医療機器）
- 医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律第 80 条の 2（治験の取扱い）第 2 項（治験の届出を要する医療機器に関する規則）、第 3 項、4 項、5 項
- 医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律施行規則第 114 条（承認申請書に添付すべき資料等）第 19 項

  一　医療機器についての承認

  へ　臨床試験の試験成績に関する資料又はこれに代替するものとして厚生労働大臣が認める資料
- 医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律施行規則第 274 条（機械器具等に係る治験の届出を要する場合）
- 「臨床研究に関する倫理指針」（平成 20 年厚生労働省告示第 415 号）※平成 27 年 4 月 1 日からは「人を対象とする医学系研究に関する倫理指針」（平成 26 年文部科学省・厚生労働省告示第 3 号 ）
- 臨床研究に関する倫理指針質疑応答集（Q&A）の改正について（医政研発第 0612001 号平成 21 年 6 月 12 日）
- 「コンピュータ診断支援装置に関する評価指標」（平成 23 年 12 月 7 日薬食機発 1207 第 1 号（別添 3））
- コンピュータ診断支援装置におけるソフトウェア設計・開発管理 開発ガイドライン

2012

- 「画像診断ワークステーションのウイルス対策ソフトに関するガイドライン（技術資料 No.JESRA TR-0035-2010 制定 2010 年 6 月 25 日）」
- 医療機器の臨床試験の実施の基準に関する省令（平成 17 年 3 月 23 日厚生労働省令第 36 号、最終改正：平成 21 年 3 月 31 日厚生労働省令第 68 号）
- 機械器具等に係る治験の計画等の届出等について（平成 19 年 7 月 9 日薬食発第 0709004 号）
- 機械器具等に係る治験の計画等の届出の取扱いについて（平成 24 年 2 月 21 日薬食機発 0221 第 1 号）
- 独立行政法人医薬品医療機器総合機構に対する機械器具等に係る治験不具合等報告について（平成 25 年 3 月 29 日薬食発 0329 第 14 号）

APPENDIX

A.1　CAD の分類と CADx の位置付け

コンピュータ検出／診断支援（Computer-Aided Detection/Diagnosis; CAD）とは、X 線画像に代表される放射線画像をはじめとする医用画像をコンピュータで定量的に解析し、「医師による診断」を側面より支援する行為や機能であり、それを提供するソフトウェアあるいはそれを具備する装置を指す。

CAD のうち、解析結果に基づいて病変候補位置の情報をマーカなどで医師に示すだけの行為や機能を有する場合はコンピュータ検出支援（Computer-Aided Detection; CADe）、質的診断に関する情報までも提示する行為や機能を有するものはコンピュータ診断支援（Computer-Aided Diagnosis; CADx）として分類する。

CADe および CADx は一般的にはソフトウェアとしてハードウェアに搭載された製品であるので、その機能を実現するシステム構成としては、ソフトウェア単体の場合と CAD ソフトウェアを搭載した装置である場合の２通りがある。

以上の観点から、CAD の分類や本ガイドラインで扱う CADx の位置付けを図 A1 に示す。

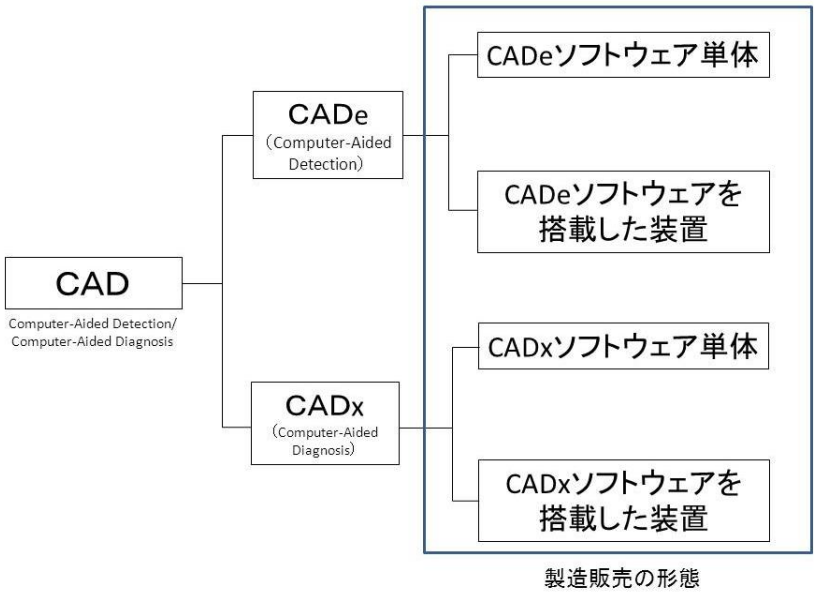

図 A1　CAD の分類と CADx の位置付け

CADx の対象となるモダリティには、主要な医用画像診断装置である X 線診断装置、X 線 CT 装置、核医学診断装置、MRI 装置、超音波診断装置などにより得られる画像のみならず、臨床で医用画像を用いて診断を行う検査として内視鏡画像、サーモグラフィー、病理顕微鏡写真なども含まれる。さらには、将来的に臨床に用いられる可能性のある光イメージン

グ装置、光超音波装置、分子イメージング装置などのモダリティも含まれる。これらのことから、単一臓器や単一疾病の良悪性の判定のみならず、多臓器・複数の疾患およびそれらの経時的な変化も対象となりえる。

A.2　性能評価法の基礎

CAD としての性能評価に対しては、以下に示す分類率や、それを用いた ROC (Receiver Operating Characteristics)解析のいずれかを用いることを推奨する。どちらを使うかは、臨床利用の観点から決定してよい。すなわち、臨床的に一組の分類率による評価が妥当な場合には分類率を用いた評価で良く、さまざまな分類率の場合を総合的に評価しなければならない場合には ROC 解析を実施するのが一般的である。

分類率や ROC の計算には下記の表（クラス分類実験結果の集計表）の数値を用いる。

表 A1　クラス分類率に基づく性能評価（クラス分類実験の集計表）

CADx の出力 / 正解(Gold standard)	クラス１	クラス２	…	クラスｎ	合計※1
クラス１	a_{11}	a_{21}	…	a_{n1}	$S_1=a_{11}+\cdots+a_{n1}$
クラス２	a_{12}	a_{22}		a_{n2}	$S_2=a_{12}+\cdots+a_{n2}$
:	:	:	:	:	:
クラスｎ	a_{1n}	a_{2n}	…	a_{nn}	$S_n=a_{1n}+\cdots+a_{nn}$
合計※2	$a_{11}+\cdots+a_{1n}$	$a_{21}+\cdots+a_{2n}$	…	$a_{n1}+\cdots+an_{1n}$	$S=S_1+\cdots+S_n$

※１正しく分類された数についての正解クラス別の合計

※２正しく分類された数についての出力別の合計

これらの数値は、CADx の実際の利用を想定して作成したデータベースを用いて、Leave one out 法、k-fold Cross Validation 法、Hold Out 法のいずれかによって求めることが望ましい（Appendix「A.6 性能評価手法」参照）。ここで、評価に用いたデータベースが実際の利用を想定しているためには、ランダム標本か階層標本など、学術的に妥当な方法を用いてデータ収集を行えば良い（Appendix「A.4 画像データベースとゴールド・スタンダード」参照）。

また、正解は病理診断結果や合理的な診断結果から導かれていなければならない。学術的に裏付けのないクラスを独自に設定して CADx の性能評価に用いても，裏付けのない性能評価にしかならない。ここで、「合理的」とは、学術的な客観性や再現性があることを指す。従って、手術や生検で得られた病理所見、画像検査で決定した診断、経過観察などを含む画像検査で下された診断、上記の検査の結果に基づく総合的な判定などがあり得る。

ただし、様々な要因によって検査結果に大きなばらつきやバイアスが含まれる事例では、複数の医師による診断結果の平均や合意などにより差異を小さくするか、より精密な他の検査に置き換える必要がある。なお、ファントムなどの人工データを利用する場合には、妥当な方法で定義された正解を用いねばならない。

分類率による評価

正分類率は、下記のように、表 A1 中の数値から計算される。

$$\text{クラス i の正分類率} = a_{ii} / S_i \quad (i=1 \ldots n)$$

ここで、あるクラスの診断精度は従来よりも良くなったが、他のクラスの精度が悪くなったのでは意味が無いことに注意をしなければならない。十分な数のデータを用いた上で、すべてのクラスに対する精度が同等か、あるいは、精度が統計的に有意に向上したクラス（クラス名を明確にすること）が一つ以上であり、かつ、その他のクラスに関する性能は統計的に同等であることを示す必要がある。

ROC (Receiver Operating Characteristics)解析による評価

2 クラス分類の場合は各クラスの正分類率の間に存在するトレードオフの関係を表した ROC 曲線（Appendix「A.7 観察者特性の評価」参照）を用いることにより統計的に証明することができる。ここで、精度が同等であることを示す場合には、統計的に十分な数のデータに基づいていなければならない。また、ROC の軸は、上記の各クラスの正分類率に基づいて定義する。例えば、良悪性鑑別の場合には、縦軸は感度、横軸は偽陽性率となる。ただし、対象によっては、LROC（Localized response ROC）、FROC（Free response ROC）、AFROC（Alternative FROC）、および JAFROC（Jackknife AFROC）などの他の評価法が適していることがある[1)]。その場合には、上記の分類率を適切な性能指標、例えば FROC の場合には、偽陽性率を一症例あたりの偽陽性数などに置換して用いることが必要である。

良悪性の判定のみならず、鑑別診断結果をリストアップする場合や、疾患の程度の分類を行う場合などは多クラスの分類が必要となる。この場合は、ある１つの注目するクラスと、残りの(n-1)のクラスを１つにまとめたクラスとの 2 クラスについて行う ROC 解析を、注目するクラスを替えながら n 通りについて行えばよい。一部のクラス群のみに注目して ROC 解析を行う方法も考えられるが、その場合には、一部のみに注目することの臨床的妥当性や、評価の統計的妥当性を学術文献などにより確認し、根拠を明確にする必要がある。

【参考文献】

1) 尾川ら編：医用画像工学ハンドブック（Part II, §3.3～§3.5），日本医用画像工学会，2012.

統計的検定に関する注意事項

全ての検定は、統計学上の手続きを踏まえたものでなければならない。仮説検定における P 値などの有意水準としては統計的に妥当なものを用いる。読影実験の際の施設数や医師数はそれぞれ複数が望ましいが、1 施設でも良いと自己判断する場合は科学的に妥当な根拠を示すことが不可欠である（Appendix「A.8 データ収集する施設数」）。ただし、多クラスの場合にはクラス毎のデータ数に偏りがあり、クラス間のデータ数の比率が実際の臨床における症例数とかけ離れる可能性が高くなる。そのため、特に注意をして十分なデータ数および施設数を確保する必要がある。なお、クラスごとの分類率にコスト（重み）を導入して再定義し、それに基づいて統計的検定を実施する場合には、検定を実施するより前に、そのコストが臨床的・学術的に正当であることを、査読を受けた論文などで確認して根拠を明確化しなければならない。また、各クラスのデータ数については臨床適用可能な状態で統計解析の結果において有意な差を示すために十分な数を用いることが重要となる。

【参考文献】

1) Wagner et al., Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol., 14(6):723-748, 2007.

2) 藤田ら(監修), 実践医用画像解析ハンドブック（§6.3.3）, オーム社, 2012.

A.3 CAD に対する QMS (Quality Management System)

品質マネジメントシステム（Quality Management System, QMS）は、製品や提供されるサービスの品質を管理するために用いられるシステムである。CADx の場合、医療機器に該当するので、厚生労働省令第 169 号 平成 16 年 12 月 17 日 医療機器及び体外診断用医薬品の製造管理及び品質管理の基準に関する省令（QMS 省令）の適用が求められる。品質保証活動を統括する責任者として、責任技術者を置き、設計、購買、製造、検査などのプロセスや是正処置などを管理する体制構築が求められる。

医療機器の使用者および患者の安全性確保のため、製造業者あるいは製造販売業者は、設計や製造および使用の各段階で製品のリスクを適切にコントロールすることが求められる。リスクマネジメントの規格としては、ISO14971 がある。

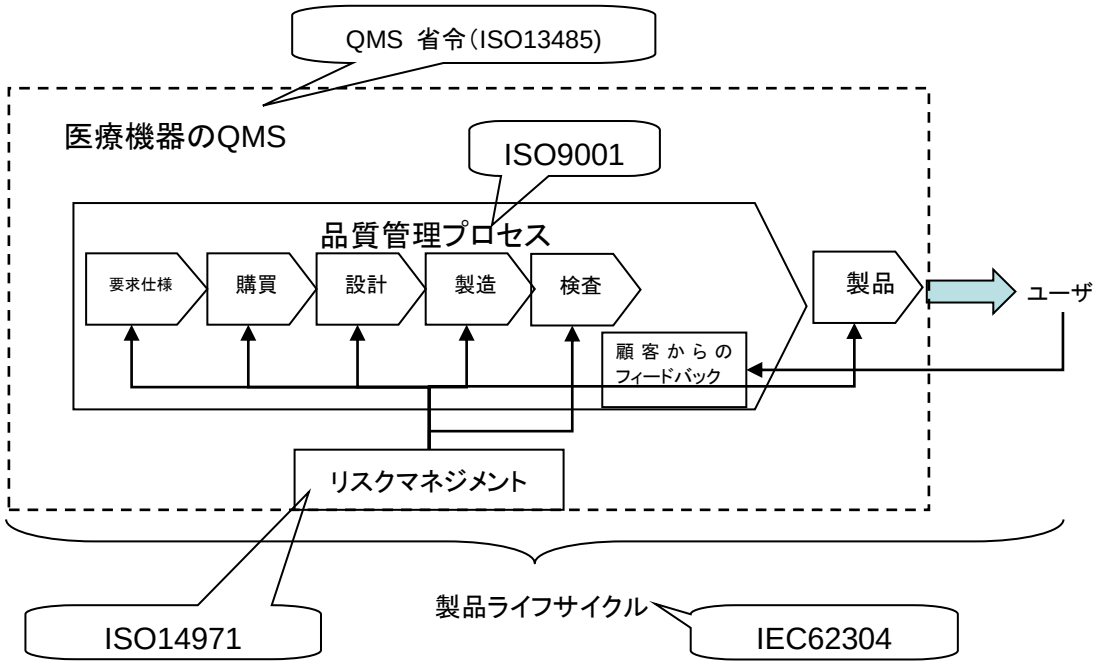

図 A2：医療機器の品質マネジメントシステム（Quality Management System, QMS）

（開発 WG 委員会からの提案）

また、ソフトウェア製品の場合、QMS のみならず、IEC 62304 の適用が近年要求される状況となっている。特に、欧米では、IEC62304 は、認知規格（Recognized Standards）として､製造業者あるいは製造販売業者に適用が求められている｡IEC 62304 はソフトウェア開発プロセスに重点を置き、ソフトウェア開発と検証活動に適用される規格である。具体的には、ソフトウェア開発計画や要件分析、アーキテクチャ設計、ソフトウェア設計、ユニット実装と検証、ソフトウェア統合と統合試験、システムテスト、そして最後にソフトウェアリリースなどの活動に対して要求事項が規定されている。日本においても、製品開発に際しては JIS T 2304 や IEC/TR80002-1「医療機器ソフトウェアへの ISO14971 の適用の手引き」を適用することが望ましい。

以上の状況を概念的にまとめると図A2のようにQMSおよび関連規定を適用することを推奨する。これらを加味した「コンピュータ診断支援装置におけるソフトウェア設計・開発管理開発ガイドライン 2012」も策定されているので、同時に参照されたい。

A.4　画像データベースとゴールド・スタンダード

CAD 研究において、CAD システムのトレーニング（または学習）およびテストに使用される画像データベースの特性は非常に重要であり、その特性によって CAD 研究の結果が大きく左右されるといっても過言ではない[1)]。CAD 研究に用いられる画像データベースには、多くの場合、画像だけではなく、様々な付帯情報が含まれる必要がある。その情報としては、細胞診や組織診などの病理診断により得られた確定診断結果や専門医による客観的な診断の難易度、患者の性別や年齢などが挙げられる。こういった画像データベースの構築においては大量の画像と幅広い症例の情報を収集し、それらを整理するための工夫が必要となる。しかしながら、単一の施設では収集できる症例数が限られるため、大規模な画像データベースの構築が困難な場合が多い[2)]。そのため、デジタル画像が医用の世界に普及し、CAD 研究が盛んに行われるようになった頃から、誰もが使用可能な公共の画像データベースの構築に関する研究が国内外で行われてきた[2-5)]。現在では、いくつかの公共での使用が可能な画像データベースが公開されており、それらには、日本放射線技術学会で構築された結節影あり／なしの胸部単純 X 線写真のデータベース[2)]や、特に読影学習を目的としたマンモグラフィのデータベース[3)]、米国の National Cancer Institute (NCI)の研究班が構築した肺結節の CT 画像のデータベースである Lung Image Database Consortium (LIDC)[4)]、南フロリダ大学のディジタルマンモグラフィのデータベース[5)]などがある。このような画像データベースの存在により、画像データベースの構築が困難な研究者でも量的および質的に充実した研究用データを使用し、CAD の研究を行うことが可能となり、また同じ画像データベースを使用しているものであれば、開発された CAD システムの性能の相互比較が可能となった。新しいモダリティや検査法に応じた CAD の開発においては、各研究施設において研究目的に応じた画像データベースを構築することも想定されるが、その際には構築したデータベースによって、CAD の評価がバイアスを受けることのないように注意する必要がある。

画像データベースに収録される画像の収集を計画する場合、その母集団について~~まず~~考慮する必要がある。例えば、過去数年間にある医療機関を受診したすべての患者を対象とするのか、ある特定の検査を受けた患者のすべてを対象にするのかによって、収集する画像の母集団は違ってくる。CAD の開発を行う場合に、対象となる疾患の罹患率は重要な要素となるので、受診したすべての患者を母集団とすれば、およその罹患率を推定することが可能になる。しかし、よほど罹患率の高い疾患でない限り、CAD 開発のための画像データベースを構築するために必要な母集団の数は膨大なものになると予想される。そのため、多くの場合は、CAD が対象とする疾患の疑いがあり、対象とするモダリティで検査を受けたすべての症例を母集団として設定し、その母集団からランダム標本、もしくは階層標本によって研究に必要な症例数を確保する研究デザインが採用される。ここで、ランダム標本とは、母集団からランダムに標本を抽出し、データベースに収録する方法である。他方、階層標本とは、母集団の特性（年齢・性別構成、既往歴の有無など）を調べた上で、その

特性に合わせて標本を抽出する方法である。一般に、収録される症例の数が多くるにつれて、ランダム標本と階層標本の両者を用いた場合の差異が小さくなる。

画像の収集と同時に考慮しなければいけないのは、画像に含まれる病変のゴールド・スタンダード(Gold Standard; GS)の決定である。GS は Reference Standard や Ground Truth とも呼ばれ、臨床研究において「その症例が間違いなく CAD 研究の対象となる疾患である」または「間違いなく疾患ではない」ということを証明するための証拠のことであり、GS の決定が不明であったり、明確でなかったりする場合は、CAD 研究そのものの真偽が問われることになる。CAD 研究用の画像データベースに含まれる画像の GS の決定には、(1)手術または生検で得られた組織・細胞の病理所見、(2)病理所見と臨床判断（経過観察）の組み合わせ、(3)臨床判断のみ、(4)上位の診断システムの結果（例えば、胸部単純 X 線像に対する CT 検査の所見）、(5)専門医によるコンセンサス、(6)ファントム実験やシミュレーション信号など既知のデータを用いる方法が挙げられる。悪性腫瘍の病変を対象として CAD を開発する場合には、すべての症例に関して病理所見で診断が確定していることが理想であるが、悪性が強く疑われない場合は、生検や手術なしで経過観察の臨床処置がとられる場合が多いので、病理所見と臨床判断との組み合わせで GS の決定が行われるのが一般的である。

画像データベースに収録される症例の数は、CAD において使用される様々な識別機構(Classifier)の性能に大きく影響する[3]。一般には、症例数が多ければ多いほど CAD の性能の正当性は高くなるが、CAD 研究に必要な多くの症例を確保することは時として非常に困難であるので、限られた症例数で信頼性の高い CAD を開発するための工夫が必要となる。

【参考文献】

1) Nishikawa RM, Giger ML, Doi K, Metz CE, Yin F-F, Vyborny CJ, Schmidt RA: Med Phys, 21(2), 265-269, 1994.
2) Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, Doi K: AJR Am J Roentgenol, 174(1), 71-74, 2000.
3) Chan H-P, Sahiner B: Med Phys, 26(12), 2654-2668, 1999.
4) Li Q, Doi K: Med Phys, 34(3), 871-876, 2007.
5) ICRU Report 79. Receiver Operating Characteristic Analysis in Medical Imaging. Oxford University Press, Oxford, UK, 2008.

A.5　性能評価のための読影実験における注意点

4.2(ア)(4)の要件を証明するための読影実験ではバイアスなどが含まれているため、結果の解釈を誤ることを防ぐための代表的な注意点や、実験の再現性を担保するための注意点について述べる。その他の注意点については Appendix「A.8 データ収集する施設数」、「A.9 性能評価に対して収集しなければならないデータ数」、「A.10 入力する画像診断装置における収集条件の明確化」を参照のこと。

- 評価結果にバイアスが混入する恐れのある以下のような読影実験は避けなければならない。ただし、バイアスを混入させる因子はこれら以外にもあるので注意が必要である。
  - 同一医師群が同一症例群を用いて、CADx を利用しない場合と利用した場合の 2 回の実験を短期間で行ってしまうと、症例に対する記憶が 2 回目の読影結果に影響を与える恐れがある。
  - 読影の際の画像の提示順序は実際の臨床の場合と同様、原則としてランダムでなければならず、恣意的に決めてはならない。
- CADx の利用により想定される不利益を全て記録し、その妥当性について評価する必要がある。例えば、CADx の出力の待ち時間や CADx の出力を参照することで増加する読影時間などである。これらを測定し、有効性を上回る不利益が無いことを確認しなければならない。
- CADx を用いた読影実験の再現性についても注意を払わなければならない。すなわち、第三者による追試によって同等の性能が得られるよう、再現に必要な全ての実験条件を記録しなければならない。以下は条件の例を示すが、これ以外にも必要であれば追加し、CADx の性能を提示する際には必要なすべての条件を開示できるように準備しておくべきである。
  - 入力画像の仕様 : 画素サイズや濃度レベル数など（3 次元画像であればスライス厚やスライス間隔などのパラメータも含む）
  - 画像の撮影条件 : 撮影装置名、撮影範囲、管電圧や管電流などの撮影パラメータ、グリッドなどの器具の使用の有無、造影条件など（3 次元画像であれば画像再構成法や寝台移動速度などのパラメータも含む。フィルムをデジタル化した場合には、デジタル化のためのスキャナの仕様（階調、分解能、画像ひずみ）も明記する）
  - 対象疾病 : 疾病の種類や診断の難易度
  - 被検者情報 : 年齢、性別。必要に応じて過去の疾病や手術などの既往歴、体型
  - データ収集法 : 施設名、収集時期、画像枚数、および具体的なデータのサンプリング法（ランダム標本化や階層標本化）
  - CADx を動作させたコンピュータ環境 : CPU の性能、メモリサイズ、ディスプレイの解像度や γ 特性
  - CADx の処理パラメータ : 製品化後にユーザが変更可能な処理パラメータ。例えば、良悪性鑑別の場合は悪性度に対する閾値
  - CADx の利用形態 : second reader と concurrent reader の区別
  - CADx を利用した医師に関する情報 : 専門分野、読影経験年数、CADx の利用

法に関する事前説明、当該 CADx に関する習熟度、その他の CADx の利用経験

A.6 性能評価手法

CAD で使用される処理手法の多くは、原画像から候補領域を抽出する処理と、抽出された候補領域を良性・悪性または真陽性・偽陽性に識別する処理の 2 つに大別される。識別のための機構としては、線形判別分析(Linear Discriminant Analysis: LDA)、人工ニューラルネットワーク(Artificial Neural Network: ANN)、サポートベクターマシン(Support Vector Machine: SVM)などが含まれる。理想的には、候補領域の抽出と識別の両方の処理手法において、閾値の設定や手法のトレーニングのために用いられる画像データベースと、その処理手法をテストするための画像データベースの両方が用意されていることが望ましい。しかしながら、治験・臨床試験ではなく CAD 開発の段階において、必要な条件を備えた症例数を確保することは困難な状況も想定される。そこで、限られた画像データベースを有効に利用し、かつ処理手法の性能を正確に評価するために、以下に示す 4 つの評価手法を選択して用いることを推奨する。

1) 繰り返し代入法（Resubstitution: RB 法）

繰り返し代入法は、最も簡便で単純な方法で、画像データベースに含まれるすべての画像で、処理手法における閾値の設定や識別機構のトレーニングを行い、そして、その処理手法をテストする際にも同じ画像データベースを用いる。テストに用いられる画像がコンピュータのトレーニングに既に用いられた画像であるので、多くの場合に処理手法の性能は過大評価され、画像データベースに含まれる症例数が少ないほど、その傾向は顕著になる。CAD 開発のパイロット研究などで比較的小規模の症例数で行われる場合が多い。パイロット研究の段階である程度の性能が見込めないコンピュータ技術は、臨床的にも有用となる可能性が低く、有用なシステムを開発することが困難なことが予想されることに起因する。

2) Leave-One-Out 交差検定法（Leave-one-out cross-validation: LOO 法）

交差検定法（cross-validation）は、統計学において標本データを分割し、先ず、その一部を解析して、残る部分を最初の解析の仮説検定に用いる手法である。交差検定法では、最初に解析するデータをトレーニング用データセット、残ったデータをテスト用データセットと呼ぶ。LOO 法はラウンド・ロビン（round-robin）法とも呼ばれる交差検定法の一種で、先ず、画像データベースの中から一症例を取り出して、それをテスト用データセットとし、残りのトレーニング用データセットで学習させた処理手法のテストに用いる。その後、同じ作業をすべての症例について繰り返す。例えば、画像に 100

症例分のデータが含まれる場合、一症例を取り出して、残りの 99 症例分でトレーニングを行い、その取り出した一例をテストする処理を 100 回繰り返す。ここで、同一症例から複数の標本がデータセットに含まれている場合、上記の 1 つずつ標本を取り出す段階において、トレーニング用のデータセットに同一症例からの標本が含まれることになるため、そのことがテストにおけるバイアスになる可能性がある。したがって、同一症例からの複数の標本がデータセットに含まれる場合には、1 つの標本をテスト用として取り出すのではなく、1 つの症例からの標本のすべてをテスト用として取り出して、残りのデータセットでトレーニングを行う Leave-one case-out 法を用いる必要がある。

一般に、LOO 法は繰り返し回数が多いため非常に時間がかかる場合がある。また、LOO 法で評価される処理手法の性能は RB 法に比べて低くなる傾向があり、その差は症例数が大きくなると減少する。LOO 法は、処理手法の性能の評価という点では信頼性が高いが、処理手法における閾値の設定などが、全ての症例についてトレーニングが繰り返される毎に変化するので、臨床応用を考慮する場合には別の手法で閾値を固定させるなどの工夫が必要になる。

3) K 分割交差検定法（K-fold Cross-Validation: KCV 法）

KCV 法は交差検定法の 1 つであり、LOO 法が一症例ごとにトレーニング用データセットとテスト用データセットに分割していたのに比べて、画像データベースの全体を K 分割して、そのグループごとにトレーニング用データセットとテスト用データセットを入れ替えて評価を行う。例えば、画像に 100 症例分のデータを 5 分割して 20 例ずつのグループに分ける場合、1 つのグループ（20 例）を取り出して、残りの 4 つのグループ（80 症例分）で学習を行う。そして、その取り出した１つのグループをテストする処理を５回繰り返す。その後、得られた 5 回の結果を平均して 1 つの推定を得る。画像データベースに含まれる症例数が比較的多く、LOO 法では時間が必要となる場合に KCV 法は有用な評価法となる。また、分割数が少なくなればなるほどトレーニング用とテスト用の症例数の差が小さくなるため、独立したテストに近い結果が得られるので、臨床的な正当性が高くなる。

4) ホールドアウト検定法（Hold out method）HO 法

HO 法は KCV 法の分割数を 2 にした場合の評価法と基本的には同じであるが、HO 法では、トレーニング用とテスト用のデータセットを入れ替えない点が異なる。HO 法はトレーニング用とテスト用のデータセットが独立しているので、他の RB 法、LOO 法、KCV 法に比べてコンピュータの性能は低く評価される場合がほとんどであるが、臨床試験などで得られる結果とほぼ同等な評価を得ることが期待される。

A.7　観察者特性の評価

図 A3 に示すのは、胸部結節影の検出を目的として処理手法を開発し、その後、診断医の検出能の向上の評価のために ROC 解析が実施された対象モダリティの異なる 2 つの CAD 研究から得られた ROC 曲線である。1 つは対象モダリティが胸部単純 X 線像で、もう 1 つは CT を対象としている[1,2]。両者の ROC 曲線で、点線で示したのはどちらもコンピュータ単独の性能であり、このコンピュータの手法から得られた出力を、診断医が読影を行う際に提示しなかった場合(without CAD)と提示した場合(with CAD)の差から、開発した処理手法の有用性を検討した。この両者においては、モダリティが違う上に、処理手法の開発に用いられた画像データベースも違うので、単純には比較することはできない。しかし、コンピュータ単独の性能の評価と、そのコンピュータの出力を診断医に提示した場合の検出能の改善の程度には直接的な関係がないことは明らかである。コンピュータの出力を診断医が利用した場合の臨床における有用性は、観察者実験を行うことによってのみ証明することが可能である。そして、この臨床における有用性が証明されることが、「コンピュータ支援診断」研究にとってはもっとも重要なことである。さらに、図 A3 に示したデータから、コンピュータ単独の性能に対して、診断医がそのコンピュータの出力を利用した場合の検出における性能は、コンピュータ単独の性能を下回る場合（A）もあれば、上回る場合（B）もあることも分かる。

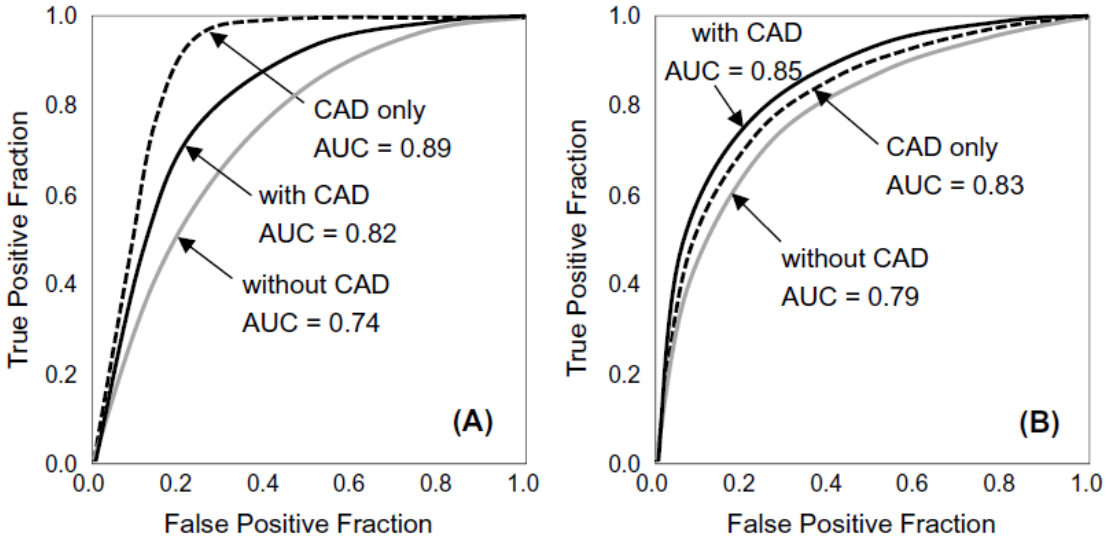

図 A3　2 種類の CAD（(A)胸部単純 X 線像における結節影の良悪性鑑別[1]，(B)胸部 CT における結節影の良悪性鑑別[2]）を使用した場合の読影医の診断能の向上の ROC 曲線

観察者特性の評価において、その結果にバイアスが含まれる可能性のある主要な因子を以下に示す。(1) 試料画像の読影の難易度、(2) 読影を行う試料画像の枚数、(3) 読影を行う試料画像の選択方法、(4) 読影実験のトレーニングに用いられる試料画像、(5) 読影実験

に使用される評価の方法、(6) 読影順序効果、(7) 読影実験に使用される環境、(8) 読影を行う観察者（読影医など）の数、(9) 読影を行う観察者（読影医）の臨床経験、(10) 実験結果に対して行われた統計的解析の手法である。これらのうち、CAD における観察者の特性の評価用データベースとして、特に考慮すべき項目は、(1)、(2)、(3)の読影試料である。

観察者実験に使用される試料画像の読影が非常に困難な場合、CAD の支援によって異常陰影が検出されていたとしても、読影医がその情報に同意しなければ読影医の性能は低くなり、CAD の支援による利益は非常に小さくなる[3]。一方、読影の難易度が非常に低い場合も、CAD の支援がなくても読影医は容易に異常陰影を検出することが可能なので、CAD の支援による利益は小さくなる。

読影に用いられる試料画像の枚数は、統計解析のために必要とされる数と、観察者が読影実験によって受ける疲労または集中力の欠如、観察者実験を分割して実施する場合の弊害などを総合的に考慮して決定する必要がある。一般に、1 回の読影実験の所用時間は観察者の疲労や集中力を考慮すると 1 時間以下が望ましいので、読影枚数が多くて１回の読影ですべての試料を観察することが困難な場合は、読影試料を分割して観察者実験を行う。なお、１つの試料に１つの信号（異常陰影）という制限のある ROC や LROC（ROC-type curve for task of detection and localization）解析に比べて、1 つの試料に複数の信号が存在することを許容する FROC（free-response receiver operating characteristic）や JAFROC（Jackknife Free-Response ROC）解析では、症例数を増やさずに信号の数を増やすことができるので、画像の数と読影に必要な時間の観点から、効率の良い観察者実験の計画が期待できる。

読影実験に用いる試料の数と同様に、試料の選択方法も実験結果にバイアスを与えないための重要な因子である。試料の選択方法には、Appendix「A.4 画像データベースとゴールド・スタンダード」で述べたランダム標本と階層標本がある。原則として、観察者実験に用いられる試料に対する CAD の性能は、処理手法の開発時にテスト用の画像データベースで得られた性能と、ほぼ同等でなければいけない。例えば、処理手法の開発時の感度と特異度が共に 75%であるにも関わらず、観察者実験に用いる試料に対する CAD の性能が 90%の感度と 80%の特異度であったとすれば、その観察者実験で得られる結論は実際の CAD を過大評価している可能性が高い。しかし、将来的に達成可能なレベルの CAD の性能を仮定して観察者実験を行う研究、例えば、CAD の性能について感度がどの程度であれば、臨床的な有用性が認められるかを証明するために、仮想的に CAD の性能を高くする場合がある[4]。

【参考文献】

1) Shiraishi J, Abe H, Engelmann R, Aoyama M, MacMahon H, Doi K: Radiology 227(2), 469-474, 2003.

2) Li F, Aoyama M, Shiraishi J, Abe H, Li Q, Suzuki K, Engelmann R, Sone S, MacMahon H, Doi K: AJR Am J Roentgenol, 183(5), 1209-1215, 2004.

3) 石田隆行, 桂川茂彦, 藤田広志監修:医用画像ハンドブック, 544-558, オーム社, 2010.
4) Shiraishi J, Abe H, Engelmann R, Doi K: Acad Radiol, 10(11), 1302-1311, 2003.

A.8　データ収集する施設数

データ収集する施設の数は、収集データの客観性を保証するために 2 施設以上で収集することが望ましい。これは、単一の施設では疾患や入力装置に偏りが生じることが懸念されるためである。単一の医療機関でのみ収集する場合は複数施設で収集した場合と同一であることを科学的に示すべきである。

A.9　性能評価に対して収集しなければならないデータ数

収集すべきデータ量の確定は開発する機器の性能を評価する上で不可欠である。「コンピュータ診断支援装置に関する評価指標」（平成 23 年 12 月 7 日薬食機発 1207 第 1 号（別添 3)）において、評価試験に必要なデータ数は「装置の目的や主要評価項目などを踏まえ、検出率や偽陽性・偽陰性率算出に必要なデータ数とする」こととしている。開発および製造販売を想定する CADx に対して標榜する性能や検出する特徴、評価方法などが異なることから、収集すべきデータ量（データ量の上限値と統計的な分布）を定量的に示すことは困難であるが、当該機器において標榜する性能が統計的に明示されることが推奨される。

A.10　入力する画像診断装置における収集条件の明確化

CADx に対する製造販売承認申請において記載すべき対象画像の収集条件は、造影剤使用の有無、撮影野（FOV）の大きさと位置、空間分解能、投影像における断面内画素サイズ、スライス厚、スライス間隔、時間分解能、撮像時間と間隔である。入力となる画像の品質により、性能の評価結果が左右される。このため、性能評価する性能に影響を与える全ての因子に対して分析する必要がある。収集条件はモダリティ別に評価することが望ましい。MRI（磁気共鳴イメージング）装置の場合を下記に例示する。

・静磁場強度（1.5T／3.0T 、など）

観測される信号および再構成される MRI 画像の SN 比は MRI 装置の静磁場強度に依存して改善する。このため、用いる MRI 装置の静磁場強度を明示すること。

・Bo や B1 の空間分布

静磁場の空間的な分布 Bo および高周波磁場の空間分布 B1 は共鳴するスピンの位相に影響を与えることから、観測される信号の SN 比、緩和時間、画像濃度の空間分布、スペ

クトル分解能に変化を与える。このため、これらを明示する必要がある。

・用いた検出コイルの感度空間特性

検出コイルにはハードウエアケージ型、CP コイルなど種々の形状があり、複数のコイルを並列受信するパラレルコイルも利用される。これら全ての検出コイルは空間的に感度が異なるため、空間的な感度分布を明示する必要がある。

・撮像法と撮像条件

測定量（T1、T2、拡散計数、化学シフト、血流、潅流、組織動作、など）は測定する組織部位の磁気共鳴元素の密度および化学的な結合状態などに依存して変化する。他方、これらを撮像するための撮像法（スピンエコー法、エコープレーナー法、など）は、共鳴するスピンをリフォーカスする技術、横磁化の喪失方法、設定するフリップ角、などが異なる。また、撮像法において設定する撮像パラメータ（繰り返し時間 TR、エコー時間、フリップ角 α、積算回数、スペクトル幅、位相エンコード数、積算回数、など）、画像再構成法は、再構成される画像の空間分解能や階調に影響を与える。これらを明示する必要がある。

平成24年度　画像診断分野

コンピュータ診断支援装置開発WG委員

座長	小畑　秀文	独立行政法人　国立高等専門学校機構　理事長
	安藤　裕	放射線医学総合研究所　重粒子医科学センター病院　病院長
	加野　亜紀子	コニカミノルタエムジー(株) 経営管理本部　商品企画部　X 線商品企画室　担当課長
	早乙女　滋	富士フイルム株式会社　メディカルシステム事業部医療政策グループ 兼　ヘルスケア事業推進室　主任技師
	椎名　毅	京都大学大学院　医学研究科　教授
	清水　昭伸	東京農工大学　大学院　工学研究院　准教授
	中田　典生	東京慈恵会医科大学　放射線医学講座
	縄野　繁	国際医療福祉大学　教授
	仁木　登	徳島大学大学院　先端技術科学教育部　ソシオサイエンス研究部　教授
	藤田　広志	岐阜大学大学院　医学系研究科　知能イメージ情報分野　教授
	古川　浩	㈳日本画像医療システム工業会　法規・安全部会　部会長 東芝メディカルシステムズ株式会社　社長附
	森山　紀之	独立行政法人　国立がん研究センター　がん予防・検診研究センター長
	諸岡　直樹	㈳日本画像医療システム工業会　法規・安全部会　副部会長 法規委員長　CAD-WG 主査 ㈱島津製作所　医用機器事業部　品質保証部　課長
	横井　英人	香川大学　医学部附属病院　医療情報部　教授